UMUTeam at TASS 2020: Combining Linguistic Features and Machine-learning Models for Sentiment Classification

José Antonio García-Díaz[a], Ángela Almela[b] and Rafael Valencia-García[a]

[a]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain
[b]Facultad de Letras, Universidad de Murcia, Campus de La Merced, 30001, Murcia, Spain

Abstract

This paper describes the participation of the UMUTeam at the TASS'2020 Workshop on Sentiment Analysis, in which two tracks were proposed. The first track consists in the classification of tweets according to general sentiments of tweets written in several Spanish varieties, whereas the second task consists in a fine-grained distinction between the six basic emotions described by Ekman (2009). Our proposal is based on the usage of linguistic features alone or in combination with word-embeddings. Specifically, we test Convolutional Neural Networks and Support Vector Machines with sentence embeddings. Although our proposal did not achieve the best results, we obtained the best precision rate regarding emotion detection (Task 2) and competitive results with respect to the general sentiment classification in which tweets written in different varieties of Spanish were mixed. We consider that our proposal, despite its limitations, provides substantial benefits such as the interpretability of the results.

Keywords

Sentiment Analysis, Supervised learning, Deep Learning, Machine learning

1. Introduction

This paper describes the participation of the UMUTeam in the TASS'2020 Workshop on Sentiment Analysis (SA). SA is a trending task of Natural Language Processing (NLP) which consists in the extraction and classification of subjective opinions in texts written in natural language.

Two tasks were proposed to the participants. The first one involved extracting the subjectivity polarity from tweets written in different varieties of Spanish, including European Spanish and different varieties from Latin America. This task was divided into two sub-tasks: (1) monolingual classification, in which several training and testing splits from different Spanish varieties were considered separately: Spain (es), Costa Rica (cr), Uruguay (uy), and Mexico (mx); and (2) multi-variant classification, in which a testing dataset consisting of mixed tweets from the above Spanish varieties were mixed. In both sub-tasks the participants were required to determine if tweets express *positive*, *negative* or *neutral* sentiments. The second task consisted in a multi-class classification of emotions based on six of the basic emotions proposed by Ekman (2009)

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)
EMAIL: joseantonio.garcia8@um.es (J.A. García-Díaz); angelalm@um.es (Á. Almela); valencia@um.es (R. Valencia-García)
ORCID: 0000-0002-3651-2660 (J.A. García-Díaz); 0000-0002-1327-8410 (Á. Almela); 0000-0003-2457-1791 (R. Valencia-García)

[1]: *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise* along with *others* in order to classify those tweets which do not agree with the aforementioned labels. All datasets, regardless of the task, were unbalanced: Task 1 contained more tweets labelled as *neutral* and Task 2 contained more tweets labelled as *others*.

We participated both in Task 1 and Task 2. In a nutshell, our proposal consisted in testing the reliability of linguistic features alone or in combination with well-known machine-learning and deep-learning models. Linguistic features, in comparison with statistical approaches, ease the interpretability of the results because they are conceptually higher-level features than words. With our participation, we attempted to answer the following research questions:

- RQ1. Is the usage of linguistic features enough to compete with state-of-the-art approaches based on statistical methods?
- RQ2. Can linguistic features be combined with statistical methods in order to improve the accuracy of the results while keeping interpretability?
- RQ3. Is the reliability of linguistic features affected by different cultural background in the same language?

To answer these research questions, we extracted linguistic features from the datasets by using a self-developed tool designed from scratch for the Spanish language and which is part of the first author's doctoral thesis. These linguistic features were evaluated separately or in combination with word-embeddings with a Convolutional Neural Network (CNN) and with sentence-embeddings with Support Vector Machines (SVM).

The rest of the paper is organised as follows. In Section 2, the datasets provided to the participants are described, whereas Section 3 describes the materials and methods used in our proposal. Then, Section 4 shows the results from each model and task along with an analysis of the results. Finally, Section 5 summarises the results according to the research questions and describes further improvement actions.

2. Corpus description

The datasets involved in these tasks were provided by the organisers of the TASS workshop [2]. The corpus was compiled from Twitter in April 2019 and it contains topics from different domains, such as politics, entertainment, catastrophes, and global events, among others. The corpus contains approximately 16k tweets for Task 1, and 8.4k tweets for Task 2 [3]. For each task, the corpus was released as three datasets, namely, training, development, and testing. Table 1 shows the distribution of the datasets for each task. Note that the size of the training and development datasets of Subtask 1.2 is 0 because participants were expected to use the training and development datasets of Subtask 1.1.

It is worth noting that hashtags and mentions were replaced by the tokens HASHTAG and @USER in order to prevent that automatic classifiers could overfit their results based on wrong assumptions. Another significant fact was that the label *Neutral* also referred to those tweets where no sentiment was assigned during the manual classification of the corpus. This is a

Table 1
Corpus size distribution for each dataset of the TASS'2020 corpus

Model	Training	Development	Testing	Total
Subtask 1.1 (es)	1127	582	1707	3416
Subtask 1.1 (pe)	967	499	1465	2931
Subtask 1.1 (cr)	778	391	1167	2336
Subtask 1.1 (uy)	944	487	1429	2860
Subtask 1.1 (mx)	991	511	1501	3003
Subtask 1.2	0	0	1501	1501
Subtask 2	5886	857	1666	8409

novelty with respect to previous editions of the TASS Workshop, in which tweets had been classified as *neutral* and *none* separately.

3. Materials and methods

In this section, we describe (1) the preprocessing techniques applied to the corpus according to our proposal (see Section 3.1), (2) the linguistic features extracted (see Section 3.2), and (3) the machine-learning models used for training the sentiment classifiers (see Section 3.3).

3.1. Preprocessing stage

Our preprocessing stage involved (1) encoding each letter to its lowercase form, (2) fixing misspelling and typos by using the Aspell library[1], (3) contracting white-space characters, such as spaces, tabs or new lines, and (4) removing expressive lengthening (the intentional elongation of letters in a word to emphasise it). The normalised version of each tweet is used as input for extracting those linguistic features that are dictionary based. However, we also kept the original version of the tweet to extract certain features regarding the usage of uppercase letters (which may be indicative of shouting or emphasis) or to identify the number of misspellings among other features.

3.2. Linguistic feature extraction

For the extraction of linguistic features we used UMUTextStats [4], a self-developed linguistic tool for text analysis that is based on Linguistic Inquiry and Word Count (LIWC) [5]. LIWC is an analysis tool that counts words which belong to pre-established categories based on part-of-speech categories as well as other categories (family, sex, and death, to name but a few). One of LIWC's greatest strengths is that has been validated under different domains. For example, it has been used in opinion mining [6], in the analysis of suicide notes [7], in cyber-bullying detection [8], or satire identification [9]. Although LIWC was originally conceived for the English language, it has a translated version available for Spanish. However, in contrast with

[1] http://aspell.net/

LIWC, our proposal handles specific phenomena of the Spanish language such as grammatical gender, as well as drawing a fine-grained distinction among PoS categories, such as adjectives, adverbs, verbs, and suffixes, among others. A further strength of UMUTextStats over LIWC is that UMUTextStats allows for complex regular expressions that are helpful in order to capture complex features such as discourse markers, by means of which different arguments connected in a text can be captured. The current version of UMUTextStats captures a total of 311 different linguistic features organised in the following categories:

- **Grammatical features (GRA)**. The features within this category are organised according to the Part-of-Speech (PoS) taxonomy. It includes verbs, adjectives, adverbs, determiners, pronouns, and conjunctions, to name but a few. Moreover, this category is organised into a more fine-grained distinction than LIWC. This decision was made because Spanish, as a highly inflected language, makes use of gender and number matching rules which may indicate when users are reporting facts, desires or hypothetical events.
- **Morphological features (MOR)**. The features within this category extract information from components of the words including prefixes and suffixes (distinguishing between different types, namely denominal morphemes, deverbal morphemes, deadjectival morphemes, and evaluative suffixation by means of diminutives, augmentatives, and pejoratives). This category also contains features to match grammatical number (singular or plural).
- **Spelling and stylistic mistakes (ERR)**. In this category, we distinguish between stylistic patterns, which denote the usage of colloquial language in order to detect atypical patterns, and linguistic errors, which can either denote the low cultural level of the writers, that is to say, errors in competence, or that they have failed to properly revise their writings before publishing them, that is, errors in performance.
- **Figurative language (FIG)**. The features within this category are related to the usage of idioms, understatements, hyperboles or any other rhetorical device whose intention is to deviate the meaning of an utterance from its literal meaning [10].
- **Linguistic processes (LPR)**. This category contains stylometric features such as number of words, syllables, and sentences. It also distinguishes between different types of sentences, such as interrogatory, exclamatory or literal quotation.
- **Symbols (SYM)**. The features within this category aims to capture sentence dividers, such as spaces, colons, semicolons among other general-purpose symbols.
- **Entity and topic extraction (ENT)**. These contain a list of general topics, including *animals*, *food*, *jobs*, *clothes* or *body-parts*, in order to determine the general topics of a text. In this category, we include the usage of inclusive language, analytic thinking, achievements and failures, and risk perception.
- **Sentiments (SEN)**. The features within this category aims to determine general words and expressions related to positive and negative feelings.

3.3. Models based on word-embeddings

The state of the art regarding SA makes use of word-embeddings and deep-learning methods. Word-embeddings, compared with traditional methods such as those models based on n-grams and one-hot-vectors, have two significant advantages. On the one hand, they represent words as dense vectors rather than sparse vectors. The main idea beyond this approach is that dense vectors allow clustering words with similar meanings. On the other, word-embeddings can be initialised by applying unsupervised techniques from general purpose trained sources, such as social networks, free encyclopedias or news sites, instead of being initialised with random values. Moreover, the idea of word-embeddings could be extended from words to whole texts, in order to represent sentences of paragraphs as dense vectors [11]. In this sense, sentence-embeddings are calculated by averaging the embeddings of the words that compose it.

Our participation in the TASS-2020 workshop involved three runs. The first run (LF + WE) consisted in linguistic features trained with a Multilayer Perceptron in combination with word-embeddings trained with a Convolutional Neural Network (CNN), the second run (LF) entailed the linguistic features trained with Support Vector Machines, and the third run (LF + SE) involved combining the linguistic features with sentence embeddings.

For the first run we used the functional API of Keras [12] to create a classifier that combines the inputs from a CNN [13], composed by a Embedding Layer with Spanish pre-trained word-embeddings from fastText [14] and a multilayer perceptron for training the linguistic features. The outputs for both deep-learning models where concatenated and combined with two more deep-learning layers and the the output layer for the final prediction. The architecture diagram of the first run is shown in Figure 1. We proposed the usage of CNN in order to exploit the spatial dimension of word-embeddings because these kinds of networks are able to find common patterns of words regardless of their position in the text, which can help solve some NLP problems such as polysemy. In this run, the training dataset was used for training and the development dataset for evaluating our proposal.

For the second and the third runs, we used the Weka platform [15] to evaluate the reliability of using linguistic features separately (LF) and in combination with sentence-embeddings (LF + SE). Both runs were trained with Sequential Minimal Optimisation (SMO), a machine-learning classifier which is based on Support Vector Machines (SVMs). Specifically, we set this SVM to use a polynomial kernel in order to learn from non-lineal models. These runs were trained with the combination of the training and development datasets.

4. Results

Our team participated in all the tasks proposed in the TASS-2020 workshop. All runs were evaluated using the macro-averaged versions of F-measure (F1), Precision (P), and Recall (R). First, Table 2 contains the results of our runs for Subtask 1.1 regarding monolingual classification. For the first run, with the combination of CNN and linguistic features, the European Spanish dataset (es) achieved the best result with a macro F1-measure of 0.503311. As precision was higher than recall, we can assume that our proposal fails regarding class detection, but when this detection is successful, the results are quite reliable. However, the rest of the Spanish varieties achieved significantly worse results between a macro F1-measure of 0.322492 for the

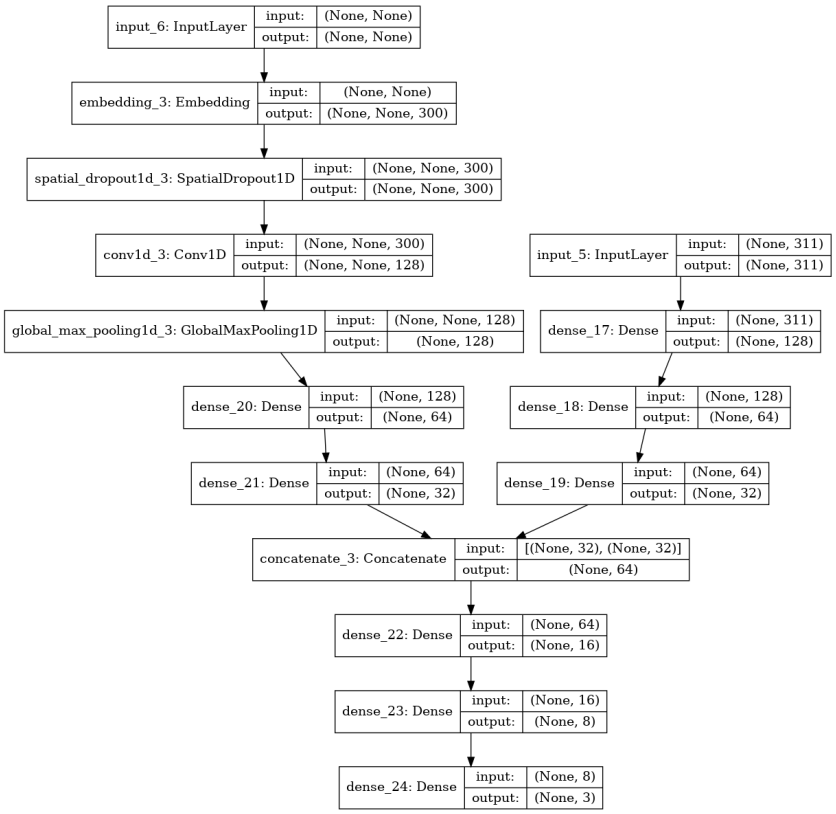

Figure 1: Model architecture of the first run, consisting in the combination of a Convolutional Neural Work for word-embeddings and a Multilayer perceptron for the linguistic features

Costa Rica dataset and 0.391288 for the Uruguay dataset. We can also observe that precision and recall were more similar between them. However, for runs 2 and 3, where the training was performed by combining the training and development datasets, the results varied and both runs achieved their best result with the Uruguay dataset (0.49832 for run 2 and 0.520168 for run 3). However, with the European Spanish dataset the results are lower than the ones achieved for the first run, owing to the lack of the development dataset during the training stage.

When our results are compared with the rest of the participants, it will be observed that they achieved more stable results and their macro-F1 measure is much more similar across the different datasets. For example, the best result, achieved by *daniel.palomino.paucar*, obtained a macro-F1 measure of 0.64667, which outperforms the results from our proposal.

Next, the results achieved by our proposal for Subtask 1.2 regarding multilingual classification are shown in Table 3. In this subtask, we merged the datasets provided for Subtask 1.1 in order to create a multilingual classifier. It will be observed in the results that the runs for this subtask were similar to the ones achieved in Subtask 1.1, achieving our best macro F1-measure with 0.357876 for run 1, but obtaining similar precision and recall rates. The best result achieved in this subtask among all the participants was 0.497966. However, as the number of participants for this subtask was low, this result must be taken with caution. It is worth noting that the rules

Table 2
Results of the UMUTeam in Task 1.1 Monolingual classification

Model	Dataset	F1	P	R
CNN (LF + WE)	es	**0.503311**	**0.561294**	**0.456186**
CNN (LF + WE)	pe	0.379346	0.399611	0.361037
CNN (LF + WE)	cr	0.322492	0.321570	0.323420
CNN (LF + WE)	uy	0.391288	0.393571	0.389031
CNN (LF + WE)	mx	0.373712	0.377036	0.370446
SMO (LF)	es	0.373223	0.373757	0.372691
SMO (LF)	pe	0.379505	0.393181	0.366748
SMO (LF)	cr	0.333147	0.333461	0.332834
SMO (LF)	uy	**0.498352**	**0.498630**	**0.498075**
SMO (LF)	mx	0.373502	0.373762	0.373241
SMO (LF + SE)	es	0.378384	0.378490	0.378278
SMO (LF + SE)	pe	0.389992	0.396927	0.383295
SMO (LF + SE)	cr	0.349812	0.349505	0.350119
SMO (LF + SE)	uy	**0.520168**	**0.518072**	**0.522281**
SMO (LF + SE)	mx	0.396592	0.397157	0.396030

Table 3
Results of the UMUTeam in Task 1.2: Multilingual classification

Model	F1	P	R
CNN (LF + WE)	0.336824	0.337451	0.336198
SMO (LF)	**0.357876**	**0.358662**	**0.357093**
SMO (LF + SE)	0.334466	0.335797	0.333146

of the tasks allowed participants to use external corpora or linguistic resources. However, our participation was limited to the datasets provided. In this sense, we could not provide a fair comparison with the rest of the participants until we analyse their approaches in detail.

Compared with purely statistical models such as the ones based on word-embeddings, linguistic features provide interpretability. It is possible, therefore, to obtain the most discriminatory linguistic features by calculating Information Gain (IG), which is a metric used by ensemble methods such as decision trees in order to determine when a new branch must be created. Figure 2 shows the 20 best features with major information for the combination of the training datasets of the different datasets provided for Subtask 1.1. We can observe that *Sentiments (SEN)* is the linguistic category which provides the largest number of linguistic features including *positive*, *negative*, *anger*, *offensive* and *sad*. Out of these features, *positive* is, by far, the most discriminatory feature. *Grammatical features (GRA)* is another linguistic feature that provides several features, such as *adverbs-negation* and *adjectives-qualifying*. It is worth noting that these linguistic features are hard to obtain by applying models based on word-embeddings like those linguistic features regarding stylistic patterns (misspellings, linguistic errors, and swear), as well as other features used to add emphasis such as the number of exclamatory sentences.

Figure 2: Information gain of the 20 best linguistic features for the Task 1.2

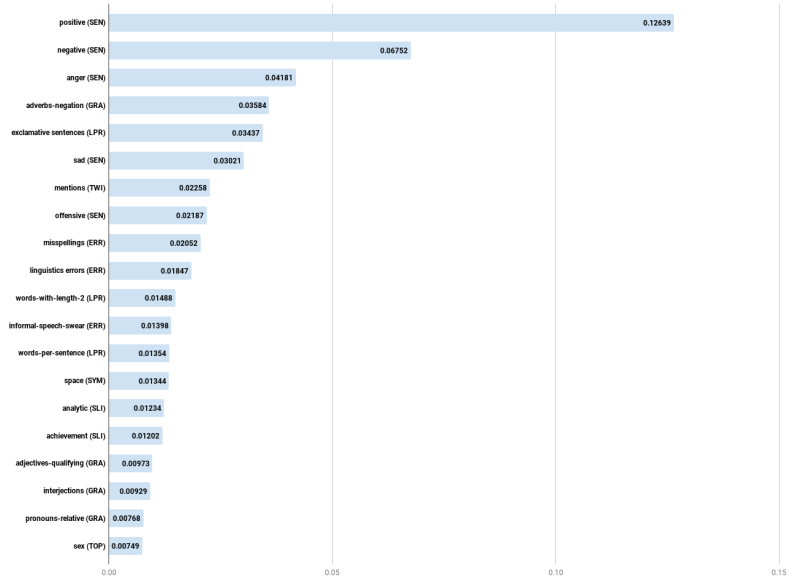

Table 4
Results of the UMUTeam in Task 2: Emotion detection

Model	F1	P	R
CNN (LF + WE)	0.242799	0.243873	0.241734
SMO (LF)	0.372774	**0.490520**	0.300614
SMO (LF + SE)	**0.378694**	0.419558	**0.345084**

Finally, the results from Task 2, emotion detection, are shown in Table 4. As can be observed, the usage of linguistic features and sentence embeddings (run 3) achieved the best result over the rest of our runs. It worth noting that run 3 only achieves slightly better results than the run 2 (LF), but with lower precision and higher recall. Our best macro F1-measure is 0.378694, which is reasonably successful, in view of the complexity of the task although the best score was achieved by *jogonba2elirf* with a macro F1-measure of 0.446582.

5. Conclusions and further work

In this paper, the participation of the UMUTeam in the TASS'2020 Workshop on Sentiment Classification applying linguistic features has been described. For Subtask 1.1, our proposal achieved decent results only for one dataset: European Spanish with CNN and linguistic features, and for the Uruguay dataset with SMO and linguistic features with average word-embeddings. For Subtask 1.2, which consisted in the combination of the datasets from different Spanish varieties, we achieved more stable results irrespective of the model applied. However, the low

number of participants hindered the achievement of better insights. Finally, for Task 2, which consisted in emotion detection, we achieved reasonably good results with a macro F1 of 0.372774 over six different classes.

As regards *RQ1* and *RQ2*, we have observed that our proposal achieves good results when comparing the usage of LF separately with the rest of our runs, but the results are far from being the best results. In this sense, we need to perform a more detailed analysis of the rest of the participants in order to identify the weakness of our proposal. Regarding *RQ3*, we have observed that our proposal does not provide stable results for Subtask 1.1 when the different dialects are considered separately, but the results were more stable when training and testing contained tweets from the different dialects as the same time. This fact suggests that some of the linguistic features between each dialect and cultural background are complementary, but not enough to perform a reliable sentiment classification.

We are well satisfied with our participation for the first time in a TASS workshop and our competition in challenging NLP tasks. We are, however, aware of the limitations of our proposal and the long way to go. To our mind, the main drawback is that we focused only on European Spanish during the design of the linguistic features. We will, therefore, focus on improvements for other varieties enabling the adaptation of the system. Regarding the technological aspect, we will include the hyper-parameter tuning in our pipeline, in order to choose the optimal hyper-parameters for a learning algorithm. We will also try to incorporate other pre-trained and contextualised word-embeddings such as ELMo [16] and BERT [17].

Acknowledgments

This work has been supported by the Spanish National Research Agency (AEI) and the European Regional Development Fund (FEDER/ERDF) through projects KBS4FIA (TIN2016-76323-R) and LaTe4PSP (PID2019-107652RB-I00). In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the Doctorado industrial programme.

References

[1] P. Ekman, Lie catching and microexpressions, The philosophy of deception 1 (2009) 5.

[2] M. García-Vega, M. C. Díaz-Galiano, M. A. García-Cumbreras, A. Montejo Ráez, S. M. Jiménez Zafra, E. Martínez-Cámara, C. A. Murillo, E. Casasola Murillo, L. Chiruzzo, D. Moctezuma, Sobrevilla, Overview of tass 2020: Introduction emotion detection, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), volume 2664 of *CEUR Workshop Proceedings*, CEUR-WS, Málaga, Spain, 2020.

[3] F. M. Plaza del Arco, C. Strapparava, L. A. Urena Lopez, M. Martin, EmoEvent: A multilingual emotion corpus based on different events, in: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 1492–1498. URL: https://www.aclweb.org/anthology/2020.lrec-1.186.

[4] J. A. García-Díaz, M. Cánovas-García, R. Valencia-García, Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious

diseases in latin america, Future Generation Computer Systems 112 (2020) 614–657. doi:https://doi.org/10.1016/j.future.2020.06.019.

[5] Y. R. Tausczik, J. W. Pennebaker, The psychological meaning of words: Liwc and computerized text analysis methods, Journal of language and social psychology 29 (2010) 24–54.

[6] M. del Pilar Salas-Zárate, E. López-López, R. Valencia-García, N. Aussenac-Gilles, Á. Almela, G. Alor-Hernández, A study on LIWC categories for opinion mining in spanish reviews, J. Inf. Sci. 40 (2014) 749–760. URL: https://doi.org/10.1177/0165551514547842. doi:10.1177/0165551514547842.

[7] B. O'dea, M. E. Larsen, P. J. Batterham, A. L. Calear, H. Christensen, A linguistic analysis of suicide-related twitter posts., Crisis: The Journal of Crisis Intervention and Suicide Prevention 38 (2017) 319.

[8] V. Singh, S. Ghosh, C. Jose, Toward multimodal cyberbullying detection, in: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, volume Part F127655, 2017, pp. 2090–2099. doi:10.1145/3027063.3053169.

[9] M. del Pilar Salas-Zárate, M. A. Paredes-Valverde, M. Á. Rodríguez-García, R. Valencia-García, G. Alor-Hernández, Automatic detection of satire in twitter: A psycholinguistic-based approach, Knowl. Based Syst. 128 (2017) 20–33. URL: https://doi.org/10.1016/j.knosys.2017.04.009. doi:10.1016/j.knosys.2017.04.009.

[10] M. del Pilar Salas-Zárate, G. Alor-Hernández, J. L. Sánchez-Cervantes, M. A. Paredes-Valverde, J. L. García-Alcaraz, R. Valencia-García, Review of english literature on figurative language applied to social networks, Knowl. Inf. Syst. 62 (2020) 2105–2137. URL: https://doi.org/10.1007/s10115-019-01425-3. doi:10.1007/s10115-019-01425-3.

[11] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al., Universal sentence encoder, arXiv preprint arXiv:1803.11175 (2018).

[12] F. Chollet, et al., Keras (2015), 2017.

[13] Y. Kim, Convolutional neural networks for sentence classification, CoRR abs/1408.5882 (2014). URL: http://arxiv.org/abs/1408.5882. arXiv:1408.5882.

[14] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, CoRR abs/1802.06893 (2018). URL: http://arxiv.org/abs/1802.06893. arXiv:1802.06893.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: an update, ACM SIGKDD explorations newsletter 11 (2009) 10–18.

[16] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, CoRR abs/1802.05365 (2018). URL: http://arxiv.org/abs/1802.05365. arXiv:1802.05365.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. arXiv:1810.04805.